Technical Report 872

AD-A219 917

Power Analysis of Gunnery Performance Measures: Differences Between Means of Two Independent Groups

John E. Morrison
Human Resources Research Organization

January 1990

United States Army Research Institute for the Behavioral and Social Sciences

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

JON W. BLADES
COL, IN
Commanding

Research accomplished under contract
for the Department of the Army

Human Resources Research Organization

Technical review by

David W. Bessemer
Bob G. Witmer

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POX, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS --
2a. SECURITY CLASSIFICATION AUTHORITY --	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE --	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) SP-PRD-88-10	5. MONITORING ORGANIZATION REPORT NUMBER(S) ARI Technical Report 872
6a. NAME OF PERFORMING ORGANIZATION Human Resources Research Organization 6b. OFFICE SYMBOL (If applicable) --	7a. NAME OF MONITORING ORGANIZATION U.S. Army Research Institute Fort Knox Field Unit
6c. ADDRESS (City, State, and ZIP Code) 1100 S. Washington Street Alexandria, VA 22314	7b. ADDRESS (City, State, and ZIP Code) Steele Hall Fort Knox, KY 40121
8a. NAME OF FUNDING/SPONSORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences 8b. OFFICE SYMBOL (If applicable) PERI-I	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER MDA903-86-C-0335
8c. ADDRESS (City, State, and ZIP Code) 5001 Eisenhower Avenue Alexandria, VA 22333-5600	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO.: 63743A; PROJECT NO.: 794; TASK NO.: 331; WORK UNIT ACCESSION NO.: H1

11. TITLE (Include Security Classification)
Power Analysis of Gunnery Performance Measures: Differences Between Means of Two Independent Groups

12. PERSONAL AUTHOR(S)
Morrison, John E. (HumRRO)

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Interim	FROM 87/12 TO 88/03	1990, January	

16. SUPPLEMENTARY NOTATION
Appendix Tables C-1 and C-2 from Introductory Statistics for the Behavioral Sciences, Third Edition, by Joan Welkowitz, Robert B. Ewan, and Jacob Cohen, (Continued)

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Statistical Power Analysis; Table VIII; Gunnery Performance; U-COFT

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Determination of sample size is a problem that has both practical and statistical implications for gunnery performance research. This report examines these implications in the context of statistical tests that compare means from two independent samples of armor crews. Performance variability estimates of gunnery data were derived from Table VIII qualifications at Grafenwöhr and from published research on the Unit-Conduct of Fire Trainer (U-COFT). These data were used in examples to describe power analysis procedures developed by Welkowitz, Ewan, and Cohen (1982) for determining power and sample size and to calculate minimum detectable differences (MDDs) between two samples of crews assuming a two-tailed test of significance with a standard significance criterion of .05 and power set to .80. One of the more notable findings from this analysis was that statistical comparisons of company-sized samples of crews (i.e., N = 14) are relatively insensitive to mean differences in speed and accuracy of performance. The limitations of the proposed methods for other tests of significance are also discussed.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL John A. Boldovici	22b. TELEPHONE (Include Area Code) (502) 624-6928 22c. OFFICE SYMBOL PERI-IK

DD Form 1473, JUN 86 *Previous editions are obsolete.* SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

ARI Technical Report 872

16. SUPPLEMENTARY NOTATION (Continued)

 John A. Boldovici, Contracting Officer's Representative.

Technical Report 872

Power Analysis of Gunnery Performance Measures: Differences Between Means of Two Independent Groups

John E. Morrison
Human Resources Research Organization

Field Unit at Fort Knox, Kentucky
Donald F. Haggard, Chief

Training Research Laboratory
Jack H. Hiller, Director

U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

January 1990

Army Project Number
2Q263743A794

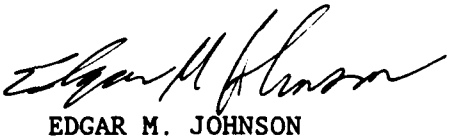
Education and Training

FOREWORD

Determination of sample size is a problem that has both practical and statistical implications for gunnery performance research. This report discusses these implications and provides a power analysis technique for calculating the minimum detectable difference (MDD) between two independent samples. This technique was used to calculate a table of MDDs for typical sample sizes found in gunnery research. Researchers can use this table to make tradeoffs between sample sizes and MDD.

This research is part of the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) task entitled "Application of Technology to Meet Armor Skills Training Needs." That task is performed under the auspices of ARI's Armor Research and Development Activity at Fort Knox, whose mission includes designing and executing human performance research in armor gunnery. The results presented in this report were briefed to the Commanding General and Staff of the U.S. Army Armor Center (USAARMC) for consideration in developing future crew and platoon qualification tables. The methods outlined in this report are being used by the Directorate of Evaluation and Standardization to determine the sample size requirements of their evaluations. Finally, the power analysis techniques were used in a companion paper entitled "Description and Prediction of Grafenwoehr M1 Tank Table VIII Performance" to determine distribution effects that are required for significant differences on Table VIII type gunnery data.

The proponent for this research is the Training and Doctrine Command (TRADOC), and the user is the USAARMC (Letter of Agreement with ARI entitled "Establishment of Training Technology Field Activity, Ft. Knox, Kentucky," dated 4 November 1983). Access to some of the data sources was provided by Mr. Al Pomey of the U.S. Army Armor and Engineer Board.



EDGAR M. JOHNSON
Technical Director

POWER ANALYSIS OF GUNNERY PERFORMANCE MEASURES: DIFFERENCES BETWEEN MEANS OF TWO INDEPENDENT GROUPS

EXECUTIVE SUMMARY

Requirement:

Determination of sample size (N) is a problem that has both practical and statistical implications for gunnery performance research. The purpose of this research was to make the techniques of power analysis more accessible for the gunnery researcher so that he can make informed decisions about sample size.

Procedure:

Performance variability estimates were obtained from gunnery performance on Table VIII qualifications at Grafenwöhr and from published research on U-COFT. These data were used in examples to describe power analysis procedures developed by Welkowitz, Ewen, and Cohen (1982) for determining power and sample size.

Findings:

Estimates of standard deviations were obtained on four measures taken from Table VIII and seven measures from U-COFT research. For the measures that were common to both media (opening time, percent first round hits, and percent hits), the estimates were remarkably similar. Using a variant of the power analysis procedures, these data were used to calculate minimum detectable differences (MDDs) between independent groups of crews using a two-tailed test of significance given the standard significance criterion of .05 and power of .80. The most notable finding from this analysis was that statistical comparisons of company-sized samples (i.e., N = 14) are insensitive to differences in speed and accuracy of gunnery performance.

Utilization of Findings:

The advantage to using the table of MDDs provided in this report is that the researcher does not have to determine a difference between means a priori. He can instead propose a performance measure and sample size and see if the value of the MDD is "reasonable" for his needs. The table also permits the researcher to make tradeoffs between sample size and detectable difference.

POWER ANALYSIS OF GUNNERY PERFORMANCE MEASURES: DIFFERENCES BETWEEN MEANS OF TWO INDEPENDENT GROUPS

CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

POWER ANALYSIS OF GUNNERY PERFORMANCE MEASURES: DIFFERENCES BETWEEN MEANS OF TWO INDEPENDENT GROUPS

INTRODUCTION

Problem

Determination of sample size (N) is a problem that has both practical and statistical implications for gunnery performance research. Samples that are too large are clearly wasteful of manpower and equipment resources. On the other hand, samples that are too small may be invalid for parametric statistical analysis. With regard to the latter point, statisticians caution that samples should be large enough that the normal distribution provides a close approximation for the sampling distribution of means. That value of N is generally regarded as 30, which is also used as a common break point between "small" and "large" samples. However, Hays (1963) stated that sample sizes as small as 10 may be large enough that the sampling distribution of means is sufficiently approximated by the normal distribution. Indeed, a casual perusal of the published research literature indicates that Ns as small as 10-12 are not uncommon.

Whereas statistical comparisons based on Ns as small as 10 may be valid in terms of the assumptions of parametric statistics, such tests may not be sensitive enough to detect meaningful differences between groups. In that regard, Boldovici (1987) elaborated on the fact that findings of no statistical differences between groups can result from causes other than the absence of actual differences between means. In examining "...the adequacy of the research and reporting upon which estimates of [training] device effectiveness are based" (p. 240), he proposed inadequate sample size as one reason that results of tank gunnery research often do not show proficiency differences due to different training conditions, and recommended that power tests be used to estimate sample sizes.

Power analyses are not typically reported in gunnery research. Perhaps these analyses are actually performed but not reported. However, it is more likely that they have not been performed at all for two reasons. First, practical power analysis procedures were first introduced by Cohen (1969) and have only recently filtered down to introductory statistical textbooks (e.g., Welkowitz, Ewen, & Cohen, 1982; Shavelson, 1988). Researchers are not likely to be as familiar with power analysis procedures as they are with older, more established statistical procedures. Second, the detailed gunnery performance data required for power analyses have not been available to researchers. However, this situation is also changing with the recent influx of data on Table VIII live-fire performance and empirical research on U-COFT simulator performance.

Research Objectives

The ultimate purpose of the present research is to make the techniques of power analysis more accessible for the gunnery researcher so that he can make informed decisions about sample size. To accomplish this purpose, the research addressed the following specific objectives:

- to present the basic concepts of power analysis in the context of gunnery research,
- to compile Table VIII and U-COFT gunnery performance data that is required to perform power analyses,
- to present some examples of how statistical power analyses can be used to test the significance of the difference between means of two independent groups, and
- to discuss the generality and limitations of the proposed power analysis techniques.

ARMOR GUNNERY RESEARCH AND THE DETERMINANTS OF POWER

To illustrate some of the fundamental concepts of power analysis, Figure 1 presents sampling distributions that apply to a significance test of the difference between means from independent groups. The two curves represent sampling distributions of the difference between measures ($\underline{M}_1 - \underline{M}_2$) under two assumptions: The left distribution assumes that the H_0 is true (i.e., $\mu_1 = \mu_2$) and is therefore centered at zero, whereas the right assumes that H_0 is false and may be centered at any value other than zero. In the present example, the actual value of μ_1 is assumed to be greater than μ_2; thus, the mean of the distribution of differences is greater than zero. On the abscissa are two values of $\underline{M}_1 - \underline{M}_2$ (i.e., -c and +c) that represent critical values of the test statistic required to reject the null hypothesis: If the obtained difference between sample means falls between c- and c+, H_0 is retained; if the differences falls outside of either criterion, H_0 is rejected. Note that in any given situation, H_0 is either true or false so that only one of the two sampling distributions actually applies. However, overlapping the distributions illustrates how the probabilities of outcomes of a statistical test are interrelated.

Two types of errors can be committed in statistical decision making. A Type I error is defined as rejecting a true null hypothesis. The probability of a Type I error is equal to α. In a two-tailed test as illustrated in Figure 1, α is divided equally between the two tails of the sampling distribution that assumes H_0 is true. A Type II error is defined as failing to reject a false null hypothesis. The probability of a Type II error (β) is represented on the distribution that assumes H_0 is false as the area that falls short (to the left of) +c. In contrast to these two errors, power is defined as the probability of making a correct

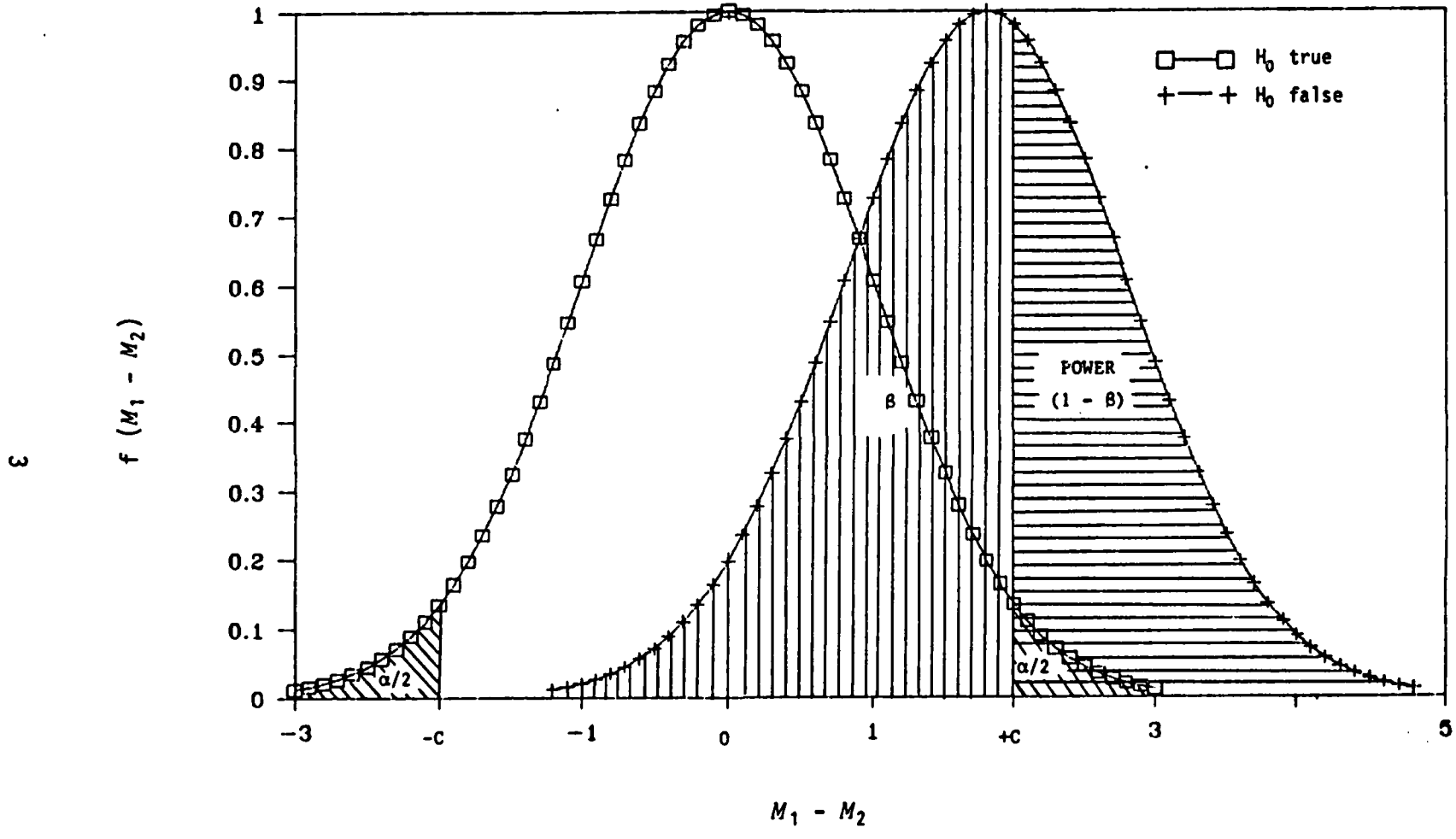


Figure 1. Sampling distribution of the difference between two means assuming H_0 is true and H_0 is false.

decision, i.e., correctly rejecting a false null hypothesis. As can be seen in the figure, power is equal to 1 - ß--the complement of the probability of committing a Type II error. In other words, power represents the sensitivity of a test to detect real differences. Thus, it is in the researcher's interest to maximize the value of power while minimizing the values of α and ß.

Power is determined by four interrelated factors: (a) the criterion of significance or α, (b) the size of the sample or N, (c) the variability of performance measures as indicated by the standard deviation or σ, and (d) the actual difference between population means or $\mu_1 - \mu_2$. The first two factors are under the direct control of the experimenter, whereas the second two are, at most, only indirectly controllable. The extent to which these factors may be controlled to affect power is discussed below with regard to standard research practices, practical constraints that face the gunnery researcher, and available gunnery performance data.

Significance Criterion

Value of α. Assuming the H_0 is true, the sampling distribution of the mean difference and α may be specified a priori. Choosing a larger (less stringent) value for α increases the power of the test. With reference to Figure 1, increasing α results in decreasing the absolute values of the test statistic required for significance ($|\pm c|$). The proportion of the right-hand curve beyond the critical value (i.e., 1 - ß) would be thereby increased. However, the price to pay for increasing α is, by definition, increasing the probability of committing a Type I error (rejecting a true H_0).

Researchers have typically set a standard value for the significance criterion at $\alpha = .05$ (two-tailed), a convention that is usually traced to Fisher's original (1925) text on analysis of variance. Statistics textbook authors often characterize the .05 level as an arbitrary convention. In contrast, Cowles and Davis (1982) argued that there are historical precedents for this value that predate Fisher's work. Furthermore, these researchers cite their own data on subjective probability suggesting that the human attribution of cause (as opposed to chance) for probabilistic events occurs somewhere between .10 and .01, a finding that supports the .05 convention. Thus, the following power analyses assume the standard .05 value for α for the sake of analytic conventions, historical precedents, and agreement with human judgment.

One- vs. two-tailed tests. Power can also be increased by using a one-tailed as opposed to the standard two-tailed test. In a one-tailed test, α is represented at one or the other tail of the sampling distribution instead of being split between two tails as shown in Figure 1. The advantage of the one-tailed test is that it effectively lowers the absolute value of c (thereby increasing power) without increasing the overall probability of a Type I error. On the other hand, only under exceptional conditions will a researcher in the behavioral sciences have enough information to make a directional prediction that is

appropriate for a one-tailed test. Even if he were able to make such a prediction, a result opposite from that predicted may not be inconsistent with other theoretical points of view. In fact, results that run counter to predictions may be the most useful in both a scientific and practical sense (D. W. Bessemer, personal communication, April 1988). For these and other reasons, statistical textbook authors (e.g., Kirk, 1984; Glass & Stanley, 1970) generally try to dissuade students from using the one-tailed procedure. Following that advice, the following power analyses will assume two-tailed tests of significance.

Sample Size

As implied by the central limit theorem, an increase in sample size reduces the variance of the sampling distribution. With reduced variance, the test statistic values fall, on average, closer to the mean value. Therefore, assuming a constant value for α, reduction of the sampling variance results in a lower absolute value of the test statistic required for significance (i.e., $|c|$). In reference to Figure 1, lowering this value would increase the proportion of the right-hand curve (H_0 false) that is beyond the critical value. Thus, increasing N increases power and reduces the probability of a Type II error without a necessary increase in α.

Firing the tank under normal conditions requires the coordinated efforts of four crewmen. Thus, the sampling unit is the tank crew rather than the individual soldier. The number of crews available for research is often constrained for practical or logistic reasons. One important constraint is that crews are frequently assigned to research projects as intact units (i.e., companies, battalions, or brigades). Assuming equal sample sizes, the resulting comparison groups are between experimental groups that are equal to these units, or some fraction thereof.[1] Therefore, it is useful to consider the standard Army armor units that may apply to research projects. Note that higher echelon units (division, corps, etc.) are not considered in the following discussion, because that have variable numbers of elements and are considered unrealistically large as individual samples.

---

[1]The fact that crews are assigned to experiments as intact units does not imply that all crews within a unit should be assigned to the same experimental condition within the experiment. Assignment of intact units to experimental conditions confounds between-unit differences with treatment effects. In addition, within-group variability estimates for intact groups underestimate the variability inherent in the population because they exclude between-unit differences (D. W. Bessemer, personal communication, April 1988). The researcher should instead randomly assign crews to experimental conditions regardless of their unit membership.

Platoon. The smallest armor unit is the platoon which consists of four tanks. For most measures, a sample size of four is too small to estimate population parameters because of the exceptionally large variability of the sampling distribution. Also, the sampling distributions of small samples are poorly fit by the normal distribution. Consequently, traditional parametric statistical techniques are inappropriate for platoon-sized samples.

Company. The next larger unit is the company which consists of three platoons having four tanks per platoon plus two additional tanks for the company commander and his executive officer. The total number of crews available from a company (14) represents perhaps the minimum acceptable sample size. Note, however, that with normal attrition that occurs in gunnery research (e.g., crews not showing up, equipment breaking down, etc.), the actual number of available crews from one company may be unacceptably small for parametric analysis.

Battalion. The next larger unit is the battalion, which consists of four companies having 14 tanks per company plus two additional tanks for the battalion commander and his executive officer. With its 58 total crews, the battalion would provide enough crews to fulfill the most rigorous requirement of parametric statistics even with substantial attrition.

Brigade. The largest unit under consideration is the close combat heavy brigade. According to doctrine, this type of brigade consists of two armor battalions and one mechanized infantry battalion.[2] No tanks are assigned to the mechanized infantry battalion nor are there tanks assigned to brigade headquarters and headquarters company. The resulting sample of 116 crews provides the upper limit of sample sizes under consideration and is only rarely achieved in gunnery research.

Variability of Performance Measures

Reducing the variability of performance measures affects sampling distributions in the same manner as does increasing sample size: The variability of the sampling distributions is reduced. Therefore, decreasing the sampling variance has the same effect on power as does increasing sample size: It increases power and reduces the probability of β without an increase in α.

The researcher has only limited control over the variability of performance measures. For instance, he can minimize the impact of external sources of variability such as those related to differences in test administration and scoring. In contrast, the experimenter cannot control internal sources of variability caused by inherent differences both within and among crews. For the purposes of power analysis, however,

---

[2]Actual brigades often deviate from this doctrinal definition as required by their stated mission.

he need only estimate the magnitude of these internal sources of variability. Table 1 presents point estimates along with the corresponding 95% confidence intervals for a number of gunnery performance measures obtained from Table VIII and U-COFT data. Appendix A presents formal definitions of each performance measure in Table 1. The next sections describe how the estimates were obtained, and the final section compares the results from the two performance measurement media.

Table VIII. The Office, Chief of Armor (OCOA) maintains a detailed data base on gunnery performance on Table VIII at Grafenwöhr. This data base is implemented on an IBM mainframe computer and updated periodically. Recent data from 872 M1 crews who underwent qualification sometime in the interval from November 1986 to June 1987 were transferred to an MS-DOS-based floppy diskette and analyzed using statistical analysis software for personal computers.[3] Data on four performance measures were analyzed:

Table 1

Standard Deviations Point Estimates and 95% Confidence Intervals for Gunnery Performance Measures

		Measurement Medium			
		Table VIII		U-COFT	
Measure	Units	SD	CI	SD	CI
Target ID Time	Seconds	---	---	1.6	0.8-2.9
Opening Time	Seconds	1.7	1.6-1.7	2.0	1.2-3.4
1st Round Hits	Percent	13	12-14	14	6-33
Hits	Percent	12	11-13	11	6-18
Elevation Error	Mils	---	---	0.15	0.06-0.41
Azimuth Error	Mils	---	---	0.34	0.09-1.25
Aiming Error	"Distance"	---	---	0.28	0.15-0.52
Table VIII Score	"Points"	98	93-103	---	---

[3]The data base itself was provided by Al Pomey of the U.S. Army Armor and Engineer Board. Standard deviation values were obtained from Hoffman (1988) who described other attributes of the performance measurement distributions as well. I thank both for their cooperation in obtaining the data required for the power analysis.

opening time, percent first-round hit, percent hit, and Table VIII score. The standard deviation estimates were based on the average performance by individual crews across the ten engagements on Table VIII. Confidence intervals for each standard deviation estimate were calculated using the chi-square distribution (Kirk, 1984). The point estimates and confidence intervals for the Table VIII data are presented in the first two columns of Table 1.

U-COFT. Standard deviation estimates of U-COFT gunnery performance were based on published research performed at the ARI Armor R & D Activity at Fort Knox. Appendix Table B summarizes this literature in tabular form. In contrast to Table VIII, U-COFT performance tests are not standardized; instead, they are customized in content and length to fit the purposes and constraints of particular experiments. Note that some of the summary data are based on only a few data points. Sample point estimates of standard deviations were calculated for measures for which there were at least seven data points. The variances of the six measures (of the total thirteen) that met this criterion were transformed logarithmically to approximate a normal distribution. The transformed variances were then treated as means to calculate a single point estimate and confidence interval for each measure (Box, Hunter, & Hunter, 1978). Point estimates and confidence intervals were based on means and standard deviations of the transformed variances and were weighted by sample size. These variance estimates were then retransformed by antilogarithm and converted to standard deviation values. The results are presented in the second column of Table 1.

Summary of results and comparisons across media. For the three performance measures that are common to both measurement media (opening time, 1st round hits, and hits), confidence intervals of the standard deviation estimates from the Table VIII data were much smaller than those from the U-COFT data. This result was expected given that the Table VIII data were based on more crews and were obtained under standardized testing conditions. Despite the difference in the stability of the two sets of estimates, the point estimates of the standard deviations for corresponding performance measures are nevertheless remarkably close in absolute values. For the two accuracy measures (first round hits and hits), U-COFT estimates of the standard deviations were within the 95% confidence interval of the Table VIII estimate indicating that standard deviation estimates from U-COFT data were not unlikely estimates of Table VIII standard deviations. Despite a small absolute difference between the standard deviation estimates for the third measure (opening time), the standard deviation estimate calculated from the U-COFT data fell above the upper limit of the Table VIII confidence interval. This greater variability in opening times may be due to the difficulty in acquiring targets on U-COFT that has been reported by Graham (1986) and others. Nevertheless, the standard deviation estimate of opening times from the Table VIII data fell well within the confidence interval for the U-COFT data indicating that the lower Table VIII value is not an unlikely standard estimate for the U-COFT data.

The similarity in standard deviation estimates were unexpected given the problems associated with measuring live-fire gunnery performance (e.g., Powers, McCluskey, Haggard, Boycan, & Steinheiser, 1975; Fingerman, 1978). That is, Table VIII performance was expected to be more variable than U-COFT performance due to the greater influence of external sources of variability. Two sets of factors may be responsible for the similarities in the standard deviation estimates. First, the U-COFT is designed to closely model tank weapon effects, including some of the external sources of variability such as round-to-round dispersion effects. Second, the Grafenwöhr data were collected under relatively standardized conditions. This practice reduces external variability due to differences in test administration.

Difference Between Means

Power is directly related to the actual difference that exists between population means; as this difference increases, so does power. With reference to Figure 1, an increase in the mean difference would be represented by increasing the distance between the two sampling distributions. Assuming constant α, the effect of increasing the difference between means would then be to increase the proportion of the right-hand distribution beyond the critical value. Thus, increasing the difference between means increases power and decreases β without affecting α.

The difference between means is an inherent quality of the treatment itself and is controllable by the experimenter only within limits. In general, the experimenter should ensure that the treatment effect is as large as possible so that the power of the comparisons is sufficiently large. For example, a one-hour exposure to experimental training program may not produce a sufficient mean difference when compared to a no-treatment control; two-or three-hour exposures may be needed. This is not to say that less extreme values of the independent variable should not be compared to test for the possibility of a nonmonotonic effect. In other cases, comparisons of the most extreme values of the independent variable may not make sense.[4] Nevertheless, the researcher should be aware of the implications of this factor for research design.

In terms of power analyses, the researcher must be able to provide an estimate of the true difference between means. Clearly this value is not known a priori; if it were, the test of significance would be pointless. On the other hand, the experimenter might be able to determine what this value should be. In other words, the researcher can establish a minimal difference that he thinks is meaningful both to him and to the consumers of his research. Once this value is determined, the procedure described in the next section can be used to ensure that his test is capable of

---

[4]I thank D. W. Bessemer (personal communication, April 1988) for pointing out the advantages of comparing differences among the less extreme values of the independent variable.

detecting such a difference with predetermined power if he provides the required number of subjects.

POWER ANALYSIS METHODS

A number of different simple methods for power analysis have been recently developed (e.g., Friedman, 1982; Shavelson, 1988; Welkowitz, Ewen, & Cohen, 1982), each algebraically equivalent to the other. However, the procedure outlined by Welkowitz et al. (1982) is notable for its simplicity and clarity. A central concept in their technique is effect size (γ), which effectively combines two determinants of power: the true difference between means and the variability of performance measures or

$$\gamma = (\mu_1 - \mu_2)/\sigma. \qquad (1)$$

Welkowitz et al. (1982) use the concept of effect size to partition power analysis into three components: γ, $\underline{N}$, and power. Specification of any two of these components necessarily determines the third.

The following subsections describe two general power analysis problems that are discussed by Welkowitz et al. and a third approach that is specifically tailored to armor applications. In each problem, it is assumed that the means from two independent groups are being compared. The generality of this design to other, more complicated designs is discussed in the final section. Sample size ($\underline{N}$) refers to the number of crews assigned to each group. Assuming equal sample sizes (i.e., $\underline{N} = \underline{N}_1 = \underline{N}_2$), the total number of crews assigned to such an experiment would be equal to $2\underline{N}$. If samples are not equal, the value of $\underline{N}$ is calculated as the harmonic mean of the two sample sizes or $2\underline{N}_1\underline{N}_2/(\underline{N}_1 + \underline{N}_2)$.

Two General Power Analysis Problems

Welkowitz et al. (1982) describe two types of power analysis problems that may potentially interest the gunnery researcher: power determination and sample size determination. Each of these is described below along with examples of armor gunnery performance problems.

Power determination. The power of a test can be calculated either before or after the fact provided the researcher has the following data: (a) an estimate of the actual difference between means, (b) an estimate of the standard deviation of performance measures, and (c) a proposed or actual sample size. To calculate power, the researcher must first combine the mean difference and the standard deviation into an effect size measure using Formula 1. The value of γ and $\underline{N}$ are then used to calculate δ (delta) as follows:

$$\delta = \gamma\ (\underline{N}/2)^{1/2}. \qquad (2)$$

Power is a direct function of δ and can be simply obtained from Appendix Table C-1 (from Welkowitz et al., 1982).

As an example problem, suppose a researcher suspects that a new training program would, at most, decrease average opening time on U-COFT by about one second. Furthermore, he knows that he can obtain only two companies of crews for his research. Thus, comparisons between the two companies would be based on a sample size of 14. With this information, he can calculate the probability of detecting a true difference before performing the research. First, calculating effect size from Formula 1, we have $\gamma = 1.0/2.1 = 0.48$. Substituting the values for gamma and sample size in Formula 2, we obtain $\delta = 0.48(14/2)^{1/2} = 1.27$. Assuming a two-tailed test and $\alpha = .05$, the expected power of the test would be about .26 (value from Table C-1). In other words, the researcher would be able to reject the null hypothesis in about one out of four experiments given this actual difference. Because of the low power, the researcher should consider changing the design of his experiment to somehow increase the effect of training or to increase sample size.

Sample size determination. An appropriate sample size may be determined if the experimenter knows (a) the desired power level, (b) the standard deviation of the performance measure, and (c) the actual difference between means. Manipulating Formula 2 to solve for N produces the following equation:

$$N = 2(\delta/\gamma)^2 \qquad (3)$$

To continue the previous example, the researcher may conclude that the easiest way to increase the power of his statistical test is to increase his sample size. To determine an appropriate sample size, he must first decide on an "acceptable" value for power. For sake of argument, assume that the experimenter considers a test sufficiently powerful if it correctly rejects the null in two out of three cases, i.e., if power is at least .67. From Appendix Table C-2, we see the δ value corresponding to a power level of .67 is 2.39. Substituting this value into Formula 3, we obtain $N = 2(2.39/.48)^2 = 49.6$. In other words, the study would require sample sizes of at least 50 crews, or 100 crews in all. In terms of unit constraints, this result implies that each sample should consist of about one battalion's complement of crews (i.e., $N = 58$).

Determination of Minimum Detectable Difference

A third technique of power analysis is added to the two previously described techniques. This method capitalizes on the fact that some of the values of power determinants are either known or are constrained in gunnery research. This third power analysis technique may be termed determination of the minimum detectable difference (MDD) between means. The MDD is the smallest actual difference between means ($\mu_1 - \mu_2$) that can be determined to be significant given values for (a) sample size, (b) the

To obtain a formula for the minimum detectable difference, either Formula 2 or 3 may be solved for γ resulting in

$$\gamma = \delta (2/\underline{N})^{1/2}. \quad (4)$$

Then substituting the Formula for γ and solving for $\mu_1 - \mu_2$, the resulting formula for MDD is

$$\mu_1 - \mu_2 = \sigma\delta(2/\underline{N})^{1/2}. \quad (5)$$

The first parameter for this analysis (sample size) is constrained to a few likely values, i.e., the numbers of crews in companies, battalions, and brigades. The second parameter (standard deviation of performance measures) may be obtained from empirical data sources. The third parameter (desired power level) may be set according to the following statistical convention. Researchers regard the consequences of a Type II error (failing to detect an actual difference) as less serious than the consequences of a Type I error (detecting a difference that is not real). Some have suggested that a ratio of 4 to 1 (Type II to Type I error probabilities) is an acceptable relationship between the two types of error (Kirk, 1984). Using this reasoning, the .05 level for α implies that .20 is an acceptable value for β. Because β is the complement of power (1 - β), the commonly accepted value for power is then .80. In terms of Equation 5, power of .80 implies a δ equal to 2.8 (from Appendix Table C-2).

Given that all the parameters of Equation 5 may be specified, a table of minimum differences for each performance measure may be generated assuming a two-tailed test and $\alpha = .05$. Table 2 shows MDD values for each of the two measurement media: Table VIII and U-COFT. Comparing across measurement media, the corresponding values for MDD are similar owing to the nearly equivalent standard deviation values shown in the previous table. Perhaps the most notable generalization that may be drawn from this analysis is that tests comparing company-sized samples are relatively insensitive to differences between means. In order to be detected by statistical test in 8 out of 10 cases, actual mean differences from company-sized samples ($\underline{N} = 14$) must on the order of 2 seconds in opening time, over 12% in hit probability, and over 100 points in Table VIII score. Furthermore, Hoffman (1988) showed that average performance on these measures for the Table VIII data is already near the limit of performance.[5] Ceiling and floor effects make it extremely unlikely that treatments can improve average performance enough to be detected by statistical comparisons of two groups. The conclusion drawn from these data is that, whereas company-sized samples may be sufficient to fulfill the requirements of parametric statistics, they are insufficient to detect all but the most drastic differences in gunnery performance.

---

[5]His Table VIII data indicate that crews average 2.1 seconds in opening time, 81% hits, and 845 Table VIII points (out of a possible 1000).

Table 2

Minimum Detectable Differences for Gunnery Performance Measures Obtained on U-COFT or on Table VIII Assuming $\alpha = .05$ and Power (i.e., $1 - \beta$) = .80

Performance Measure Sample Size[a]	Units	Medium	
		Table VIII	U-COFT
Target ID Time	Seconds		
Company		---	1.7
Battalion		---	0.8
Brigade		---	0.6
Opening Time	Seconds		
Company		1.8	2.1
Battalion		0.9	1.0
Brigade		0.6	0.7
First Round Hits	Percent		
Company		14	15
Battalion		7	7
Brigade		5	5
Hits	Percent		
Company		13	12
Battalion		6	6
Brigade		4	4
Elevation Error	Mils		
Company		---	.16
Battalion		---	.08
Brigade		---	.06
Azimuth Error	Mils		
Company		---	.36
Battalion		---	.18
Brigade		---	.13
Aiming Error	"Distance"		
Company		---	.30
Battalion		---	.15
Brigade		---	.10
Table VIII Score	"Points"		
Company		104	---
Battalion		51	---
Brigade		36	---

[a]Sample sizes are 14, 58, and 116 for company, battalion, and brigade respectively.

The advantage to using this table of minimum detectable differences is that the researcher does not have to determine a difference between means a priori. He can instead propose a performance measure and sample size and see if the value of the MDD is "reasonable" for his needs. In other words, this analysis requires that the researcher confirm a difference value from a table rather than estimate such a value. This table also permits the researcher to make tradeoffs between sample size and detectable difference. Nevertheless, the table of MDDs should not be regarded as a table of immutable values. The table should instead be regarded as a best guess at the relationship between the two factors. Other specific limitations of this approach to power analysis are discussed in the next section.

LIMITATIONS OF THE METHODS

Although these techniques should help the gunnery researcher to make more systematic decisions about sample size, there are situations that may invalidate (or at the least, limit) the interpretation of the results from the present approaches. For instance, these techniques apply to statistical hypotheses about means and not to hypotheses about other attributes of performance distributions (i.e., the correlation coefficient, r). However, analogous procedures could easily be developed for such attributes. Other less obvious boundary conditions and their effects on power are discussed in the following paragraphs.

Sample Sizes Other Than Those Specified

Although sample sizes are constrained by the organization of armor units, samples sizes other than 14, 58, and 116 are not only possible but likely under some circumstances. For instance, the tanks of the company and battalion commander and their executive officer may not be available to the researcher. Under that assumption, one would obtain unit sizes of 12, 48, and 96 for company, battalion, and brigade separately. The MDDs for these sample sizes should be slightly larger than the tabled values for corresponding units. For instance, tabled values for the hits measure on the U-COFT are 12, 6, and 4 for company, battalion, and brigade, respectively. Assuming that commanders and executive officer tanks are not available, the values of MDD recalculated from Formula 5 would be 13, 6, and 4--not much difference. As a second example, different sample sizes could be obtained by concatenating or dividing units. For instance, one could design an experiment that divides a battalion in two groups, each group consisting of two companies' worth of tanks, i.e., $N = 28$. In either case, one could determine the MDD for these particular sample sizes by using Formula 5. For instance, assuming one would want to use samples consisting of two companies, the MDD for percent hit would be $(11)(2.8)(2/28)^{1/2} = 8.2$. An even simpler procedure is to recognize that a sample size of two companies falls about halfway between one company and one battalion. Thus, one would estimate that the MDD also falls about midway between the two points or $(6 + 11)/2 = 8.5$--again, not far from the

actual calculated value. Thus, the table provides enough data points so that the MDDs of sample sizes not listed on the table may be estimated or interpolated.

Comparisons Among More Than Two Groups

Strictly speaking, the present techniques do not apply to experimental designs that compare more than two groups, i.e., those requiring one-way analysis of variance (ANOVA) techniques. Hinkle and Oliver (1983) provided methods for power analysis and sample size determination for comparisons of more than two groups. These researchers also demonstrated by example how the technique can be extended to determining the sample size requirements for the main effects of higher order designs. They acknowledged, however, that determining the sample size for testing interaction effects would be much more complex. A more serious problem with this technique is that it is based on the differences between the means of the two most extreme groups and assumes that the remaining groups do not differ from the grand mean. If the intervening group means take values other than the grand mean, the estimate of the between-groups mean square will necessarily be larger (Kirk, 1968). As a consequence, power estimates for more than two groups tend to under-estimate actual values, and sample size estimates overestimate actual requirements.

In many research projects entailing more than two groups, the analysis nevertheless focuses on comparisons between two means at a time. If the researcher were able to specify a meaningful set of orthogonal comparisons between means a priori, the procedures described herein should apply for each comparison. The rationale for this assertion is that the α for each orthogonal comparison is equal to the stated experiment-wise significance criterion. That is, the stated relationships among α, β, N, and MDD as given in Table 2 should apply to each orthogonal comparison. If the comparisons are not orthogonal, the probability of a Type I error for each comparison is greater than α, the experiment-wise error rate. Thus, the sample size estimates would tend to underestimate sample requirements for nonorthogonal requirements. [For an extensive discussion of different approaches to correcting the error rate for nonorthogonal comparisons, see Kirk (1968).]

Within-Crew Designs

An alternative to assigning independent samples of crews to treatments is to assign a single sample to all experimental treatments. Such "within-crew" designs are more powerful than between-group designs because differences between crews can be isolated and "removed" as a source of error. The residual standard deviation may be calculated as

$$\sigma_{res} = \sigma(1 - r^2)^{1/2} \qquad (6)$$

As can be seen from Formula 6, the size of the residual standard deviation is dependent on the correlation between repeated measures across subjects:

As the correlation increases, the residual standard deviation decreases and the overall treatment effect (Formula 1) increases. [Estimates of the correlations between repeated measure of U-COFT performance are provided by Graham (1986) and DuBois (1987).] Thus, given a nonzero correlation between repeated measures, the within-crew design is more likely to detect a real difference between treatments compared to an independent groups design.

The problem with within-crew comparisons is the existence of carry-over effects between treatments. If the focus of the research is on carry-over effects per se, then the within-crew design is appropriate. For instance, a within-crew design would be appropriate if one wished to study the changes in performance as a function of skill acquisition or retention. If the research does not focus on carry-over effects, the experiment may be designed to counterbalance and actually evaluate these carry-over effects. Although such within-subject designs are potentially more powerful, they have the following drawbacks: (a) they require more complex management of the research effort to ensure that each subject gets the proper sequence of conditions, (b) they usually require more lengthy participation by each subject, and (c) they sometimes require more subjects to fill out all the sequences of conditions (D. W. Bessemer, personal communication, April 1988). Finally, even counterbalancing cannot be used to compensate for independent variables whose effects are not transitory. Consider, for instance the case where an experimenter wishes to compare the effects of two training techniques. If each crew were trained using both techniques, the effect of the treatments themselves would be irrevocably confounded with unknown facilitative and/or interfering effects between treatments. Thus, whereas the within-crew design provides a more powerful approach to testing mean differences, the design is only appropriate to a limited subset of independent variables.

Accuracy of Variability Estimates

Finally, the validity of the power analysis methods discussed in this report depends on the accuracy of variability estimates. The Table VIII data were based on a substantial number of crews and the estimates appear to be reasonably stable. Furthermore, now that Table VIII data collection is automated, these data can and should be updated from time to time. The U-COFT performance data were more problematic in that performance measures were based on fewer crews and collected under varying conditions. In order to increase the stability of these estimates that we have and to add variability estimates of new performance measures, more U-COFT variability data will be needed. This assertion only reemphasizes the importance of researchers' continuing to report estimates of performance variability along with estimates of central tendency.

REFERENCES

Abel, M. H. (1987). Effects of NBC protective equipment and degraded operational mode on tank gunnery performance (ARI Technical Report 764). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A191 233)

Black, B. A., & Abel, M. H. (1987). Review of U.S. armor crew and platoon training in preparation for the 1985 Canadian Army Trophy (CAT) competition (ARI Research Report 1442). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A185 470)

Boldovici, J. A. (1987). Measuring transfer in military settings. In S. Cormier & J. Hagman (Eds.), Transfer of learning. New York, NY: Academic Press.

Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). Statistics for experimenters: An introduction to design, data analysis, and model building. New York, NY: John Wiley & Sons.

Cohen, J. (1969). Statistical power analysis for the behavioral sciences. New York, NY: Academic Press.

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. American Psychologist, 37, 553-557.

DuBois, R. S. (1987). The M1 Unit-Conduct of Fire Trainer (U-COFT) as a tank gunnery testing device: A psychometric evaluation. Unpublished master's thesis, Western Kentucky University.

Fingerman, P. W. (1978). A preliminary investigation of weapon-system dispersion and crew marksmanship (TR-78-B5). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A077 992)

Fisher, R. A. (1925). Statistical methods for research workers. Edinburgh, Scotland: Oliver & Boyd.

Friedman, H. (1982). Simplified determinations of statistical power magnitude of effect and research sample sizes. Educational and Psychological Measurement, 42, 521-526.

Glass, G. V., & Stanley, J. C. (1970). Statistical methods in education and psychology. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Graham, S. E. (1986). The Unit-Conduct of Fire Trainer (U-COFT) as a medium for assessing gunnery proficiency: Test reliability and utility (ARI Research Report 1422). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A169 196)

Hays, W. L. (1963). Statistics. New York, NY: Holt, Rinehart and Winston, Inc.

Hinkle, D. E., & Oliver, J. D. (1983). How large should the sample be? A question with no simple answer? or... Educational and Psychological Measurement, 43, 1051-1060.

Hoffman, R. G. (1988). Grafenwöhr Tank Table VIII: Descriptive statistics (Interim report). Alexandria, VA: Human Resources Research Organization.

Kirk, R. E. (1984). Elementary statistics (2nd ed.). Monterey, CA: Brooks/Cole Publishing Company.

Kirk, R. E. (1968). Experimental design: Procedures for the behavioral sciences. Belmont, CA: Brooks/Cole Publishing Company.

Powers, T. R., McCluskey, M. R., Haggard, D. F., Boycan, G. G., & Steinheiser, F., Jr. (1975). Determination of the contribution of live firing to weapons proficiency (FR-CD(C)-75-1). Alexandria, VA: Human Resources Research Organization.

Shavelson, R. J. (1988). Statistical reasoning for the behavioral sciences (2nd ed.). Boston, MA: Allyn and Bacon, Inc.

Smith, E. P., & Graham, S. E. (1987). Validation of psychomotor and perceptual predictors of armor officer M-1 gunnery performance (ARI Technical Report 766). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A191 333)

Welkowitz, J., Ewen, R. B., & Cohen, J. (1982). Introductory statistics for the behavioral sciences (3rd ed.). New York, NY: Academic Press.

Witmer, B. G. (1988a). Device-based gunnery training and transfer between the videodisc gunnery simulator (VIGS) and the Unit Conduct of Fire Trainer (U-COFT) (ARI Technical Report 794). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A197 769)

Witmer, B. G. (1988b). Effects of degraded mode gunnery procedures on the performance of M1 tank gunners (ARI Technical Report 778). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A192 246)

APPENDIX A

DEFINITIONS OF PERFORMANCE MEASURES

Performance Measure	Definition
Target ID time	Time (in seconds) from when the target first appears to when the gunner responds "identified" to the tank commander's fire command.
Opening time	Time (in seconds) from when the target first appears to when the gunner fires the first round at the first target.
First-round hits	Percentage of total engagements wherein the gunner hits the target with the first round fired.
Hits	Total number of hits divided by total rounds fired expressed as a percentage.
Elevation error	Deviation in elevation of the reticle cross hairs from the correct aiming point expressed in mils.
Azimuth error	Deviation in azimuth of the reticle cross hairs from the correct aiming point expressed in mils.
Aiming error	Conversion of elevation and azimuth error from angular measure to a single "distance" measure calculated as (elevation error^2 + azimuth error^2)$^{1/2}$.
Table VIII score	A composite score based on performance with 10 different Table VIII engagements. On each engagement, a crew can receive a maximum of 100 points according to the number of targets hit and the time required to hit the targets. The Table VIII score is calculated by simply summing over all ten engagement scores. Procedural errors (e.g., improper fire command) can reduce this overall score.

APPENDIX B

STANDARD DEVIATION ESTIMATES FROM U-COFT RESEARCH

Reference	Experimental Condition	N	Time Measures: Open Time	Hit Time	Tg Id Time	Hit-Id Time	Accuracy Measures: % Hit (1st Rd)	% Hit	No. Ht/Ex	Elev Error	Azim Error	Aim Error	U-COFT Composites: Ret'cl Aim	Tgt Acqn	Sys Mngmt
Abel (1987)	Experiment I														
	No MOPP/NORM	12	1.84				9.42					0.24			
	Mask Only/NORM	12	2.06				4.41					0.29			
	Msk & Glvs/NORM	12	1.96				3.38					0.22			
	Msk & Glvs/EMER	12	1.69				8.56					0.26			
Abel (1987)	Experiment II														
	No MOPP/NORM	12	1.74				9.3					0.26			
	No MOPP/EMER	12	1.15				8.9					0.33			
	Msk & Glvs/NORM	12	1.97				10.26					0.26			
	Msk & Glvs/EMER	12	1.93				10.37					0.35			
Witmer (1988a)	Replication 1	12	2.08	2.82	2.6		12.66	14.02		0.45	1.37				
	Replication 2	12	1.78	2.09	1.9		11.32	12.61		0.29	0.5				
Witmer (1987b)	Fully Oper'l														
	SS/Short	12	2.35				10.58			0.14	0.1				
	SS/Long	12	2.14				8.37			0.08	0.2				
	SM/Short	12	1.97				22.68			0.14	0.7				
	SM/Long	12	2.79				17.17			0.22	0.47				
	MM/Short	12	1.93				8.99			0.1	0.44				
	MM/Long	12	2.71				16.57			0.32	0.92				
	Stab System Failure														
	SS/Short	12	2.07				6.51			0.13	0.13				
	SS/Long	12	2.64				6.51			0.13	0.14				
	SM/Short	12	2.07				10.84			0.12	0.34				
	SM/Long	12	2.75				12.76			0.22	0.9				
	MM/Short	12	2.06				16.43			0.13	0.94				
	MM/Long	12	1.95				28.83			0.17	2.46				

APPENDIX B

STANDARD DEVIATION ESTIMATES FROM U-COFT RESEARCH (Continued)

			Time Measures				Accuracy Measures						U-COFT Composites		
Reference	Experimental Condition	N	Open Time	Hit Time	Tg Id Time	Hit-Id Time	% Hit (1st Rd)	% Hit	No. Ht/Ex	Elev Error	Azim Error	Aim Error	Ret'cl Aim	Tgt Acqn	Sys Mngmt
	LRF Failure														
	SS/Short	12	2.32				12.23			0.12	0.18				
	SS/Long	12	3.26				22.31			0.17	0.7				
	SM/Short	12	2.47				10.71			0.12	0.52				
	SM/Long	12	3.18				26.16			0.42	0.86				
	MM/Short	12	2.69				13.51			0.13	0.54				
	MM/Long	12	2.65				12.43			0.36	0.79				
	Ball. Comp., GPS, LRF, & Stab Failure														
	SS/Short	12	1.38				10.25			0.1	0.08				
	SS/Long	12	1.89				24.33			1.12	0.26				
	SM/Short	12	1.57				14.77			0.14	0.41				
	SM/Long	12	1.55				17.37			0.26	1.32				
	MM/Long	12	1.63				14.97			0.07	0.61				
	MM/Long	12	2.95				21.31			0.27	0.82				
Graham (1986)	Test	32	1.5				11	11		0.17	0.31		0.31	0.38	0.15
	Retest	32	1.4				16	15		0.25	0.35		0.34	0.43	0.45
Dubois (1987)	Test	165	1.624		1.673			10.9		0.165	0.261	0.246	0.237	0.388	0.117
	Retest	165	1.692		1.511			10.1		0.117	0.248	0.242	0.232	0.288	0.099
Black & Abel (1987)	Posttest														
	Unit 2-66	7		1.34	0.62	0.89		2.84	0.3						
	Unit 3-32	4		1.12	0.49	0.56		3.33	0.29						
	Unit 3-64	3		1.56	0.28	1.39		6.06	0.33						
Smith & Graham (1987)	SM/Long/Norm	90-95	2.7				19								
	MS/Long/Norm	90-95	2.8				17								
	SM/Short/No Stab	90-95	2.6				20					0.5			
Number of Data Points			41	5	7	3	39	9	3	30	30	11	4	4	4

APPENDIX C

POWER ANALYSIS TABLES FROM WELKOWITZ ET AL.

Table C-1

Power As a Function of δ and Significance Criterion α

One-tailed test (α)					One-tailed test (α)				
.05	.025	.01	.005		.05	.025	.01	.005	
Two-tailed test (α)					Two-tailed test (α)				
δ	.10	.05	.02	.01	δ	.10	.05	.02	.01
0.0	.10*	.05*	.02	.01	2.5	.80	.71	.57	.47
0.1	.10*	.05*	.02	.01	2.6	.83	.74	.61	.51
0.2	.11*	.05	.02	.01	2.7	.85	.77	.65	.55
0.3	.12*	.06	.03	.01	2.8	.88	.80	.68	.59
0.4	.13*	.07	.03	.01	2.9	.90	.83	.72	.63
0.5	.14	.08	.03	.02	3.0	.91	.85	.75	.66
0.6	.16	.09	.04	.02	3.1	.93	.87	.78	.70
0.7	.18	.11	.05	.03	3.2	.94	.89	.81	.73
0.8	.21	.13	.06	.04	3.3	.96	.91	.83	.77
0.9	.23	.15	.08	.05	3.4	.96	.93	.86	.80
1.0	.26	.17	.09	.06	3.5	.97	.94	.88	.82
1.1	.30	.20	.11	.07	3.6	.97	.95	.90	.85
1.2	.33	.22	.13	.08	3.7	.98	.96	.92	.87
1.3	.37	.26	.15	.10	3.8	.98	.97	.93	.91
1.4	.40	.29	.18	.12	3.9	.99	.97	.94	.91
1.5	.44	.32	.20	.14	4.0	.99	.98	.95	.92
1.6	.48	.36	.23	.16	4.1	.99	.98	.96	.94
1.7	.52	.40	.27	.19	4.2	.99	.99	.97	.95
1.8	.56	.44	.30	.22	4.3	**	.99	.98	.96
1.9	.60	.48	.33	.25	4.4		.99	.98	.97
2.0	.64	.52	.37	.28	4.5		.99	.99	.97
2.1	.68	.56	.41	.32	4.6		**	.99	.98
2.2	.71	.59	.45	.35	4.7			.99	.98
2.3	.74	.63	.49	.39	4.8			.99	.99
2.4	.77	.67	.53	.43	4.9			.99	.99
					5.0			**	.99
					5.1				.99
					5.2				**

Note. From Introductory Statistics for the Behavioral Sciences (p. 363) by J. Welkowitz, R. B. Ewen, and J. Cohen, 1982, New York, NY: Academic Press. Copyright 1982 by Harcourt Brace Jovanovich, Inc. Reprinted by permission of the publisher.

* Values inaccurate for one-tailed test by more than 0.1.
** The power at and below this point is greater than .995.

Table C-2

δ As a Function of Significance Criterion (α) and Power

	One-tailed test (α)			
	.05	.025	.01	.005
	Two-tailed test (α)			
Power	.10	.05	.02	.01
.25	0.97	1.29	1.65	1.90
.50	1.64	1.96	2.33	2.58
.60	1.90	2.21	2.58	2.83
.67	2.08	2.39	2.76	3.01
.70	2.17	2.48	2.85	3.10
.75	2.32	2.63	3.00	3.25
.80	2.49	2.80	3.17	3.42
.85	2.68	3.00	3.36	3.61
.90	2.93	3.24	3.61	3.86
.95	3.29	3.60	3.97	4.22
.99	3.97	4.29	4.65	4.90
.999	4.37	5.05	5.42	5.67

Note. From Introductory Statistics for the Behavioral Sciences (p. 364) by J. Welkowitz, R. B. Ewen, and J. Cohen, 1982, New York, NY: Academic Press. Copyright 1982 by Harcourt Brace Jovanovich, Inc. Reprinted by permission of the publisher.